Ontario Hydro

RESEARCH DIVISION

LEGAL NOTICE

This report was prepared by Ontario Hydro and, while effort has been made to assure the accuracy and applicability of the material contained herein, neither Ontario Hydro nor any person acting on its behalf:

(a) makes any warranty, express or implied with respect to the use of any information, apparatus, method or process disclosed in this report or that such use may not infringe privately owned rights; or

(b) assumes any liabilities with respect to the use of, or for damages resulting from the use of, any information, apparatus, method or process disclosed in this report.

91602
new 82-10

ontario hydro
research division

COMPLIMENTARY COPY
Please Credit Ontario Hydro Research if
Reproduced in Whole or in Part

To: Dr. D. Mills
Director of Research

K-SAMPLE TEST OF HOMOGENEITY

by

Arthur H. Chan and Fernado Camacho
Reliability and Statistics Section
Operations Research Department

ABSTRACT

Tests of homogeneity are used to determine whether observations from different populations have the same distribution. For the case of two populations, the Kolmogorov-Smirnov 2-sample test statistic is usually employed. This statistic is difficult to apply to more than two populations. The Statistical Analysis System (SAS) uses an ad-hoc procedure to solve this problem, but unfortunately, the approach is incorrect. This report proposes a new *k*-sample test which is easy to apply and rectifies the problem encountered in SAS.

job	file	date	report no.
72-510-000	809.730	May 18, 1989	89-126-K

TABLE OF CONTENTS

	Page
1. INTRODUCTION	1
2. RESULTS	1
3. DISCUSSION	1
APPENDIX I	2-11
APPENDIX II	12-13
DISTRIBUTION	last page

ontario hydro
research division

To Dr. D. Mills
Director of Research

1. INTRODUCTION

Many of the projects undertaken by the Statistics Unit require testing the hypothesis of homogeneity among several populations. The currently available procedures for more than two populations, such as the Kiefer-Burke-Csörgö test and the Birnbaum-Hall test, are difficult to apply. The Statistical Analysis Systems (SAS) proposes a more practical approach, but unfortunately, it turns out to be incorrect. The objective of this report is to present a new k-sample test of homogeneity that will alleviate these problems.

2. RESULTS

The proposed k-sample test of homogeneity is based on comparisons between the empirical distribution function (EDF) of each of the sampled populations with an estimate of the hypothetical common distribution. The common distribution is estimated using a data set obtained by randomly selecting half of the observations from each of the populations. The details of the test and the proof of the results are given in Appendix I. Appendix II provides a list of definitions of terminologies used in this report. This work was accpeted for presentation at the 1989 annual meeting of the Statistical Society of Canada.

3. DISCUSSION

The proposed test is quite general. It encompass and improves some of the classic tests such as the Kolmogorov-Smirnov, the Birnbaum-Hall and the Cramer-Von Mises statistics. Furthermore, it remedies the nonparametric homogeneity tests of SAS. To make this test fully operational, it will be necessary to do further studies to determine the behavior of the power function and the sensitivity of the test to different selections of the half pooled EDFs.

Approved:

A. H. Chung
Manager
Operations Research Department

Submitted:

Arthur H. C. Chan
Mathematician
Reliability and Statistics Section
Operations Research Department

Fernando Camacho
Mathematician
Reliability and Statistics Section
Operations Research Department

AHC/FC:cmh

APPENDIX I

K-SAMPLE TEST OF HOMOGENEITY

by

Arthur H. Chan and *Fernado Camacho*
Ontario Hydro Research Division
800 Kipling Ave, Toronto, Ontario M8Z 4S5

ABSTRACT

The extension of the Kolmogorov-Smirnov 2-sample test to the k populations case has been addressed by Kiefer (1959), and by Burke and Csörgő (1976). Their test statistic for the k-sample goodness-of-fit test is based on the linear contrast of the k empirical distributions under the conditions that the sum of all the coefficients is zero. This test statistic depends on the selection of the coefficients which makes it difficult to apply in practice. The Statistical Analysis System (SAS) tries to solve this problem by using the sum of the squared differences between the empirical distribution functions (EDF) and the *pooled* one (See SAS/STAT Guide p.718). Unfortunately, this turns out to be incorrect because of the dependency among the squared differences. This paper proposes a new k-sample test which (i) extends the Kolmogorov-Smirnov, the Cramer-von Mises and the Birnbaum-Hall statistics; (ii) is easier to apply than the Kiefer-Burke-Csörgő extensions; and (iii) rectifies the problem encountered in SAS.

Key Words: k-sample test of homogeneity, Brownian bridge, Kiefer process, strong embedding.

1. INTRODUCTION

Let X_j, $j = 1,2,..,k$, be k ($k > 1$) independent random variables with distributions F_j. The classical statistical problem is to test whether or not the k distributions are identical to a common distribution, F, based on independent random samples x_{jm}, $m = 1,...,n_j$. Two tests of hypothesis can be carried out: (i) a goodness-of-fit test, when F is specified; and (ii) an homogeneity test, when F is not specified. For $k=2$, the problem of homogeneity has been solved by Kolmogorov-Smirnov (K-S) (1933 & 1939). For $k \geq 2$, Kiefer (1959) and Burke and Csörgő (1976), proposed a test based on linear contrasts of the k empirical distribution functions (EDF). Birnbaum and Hall (1960) extended the K-S statistics using the maximum contrast.

The Kiefer statistic is of the form (see Csörgő and Révész, 1981)

$$\mathfrak{I}_n = \sup_x |\mathfrak{I}_n(x)| = \sup_x |\Sigma_j\, c_j(n,x)\sqrt{n_j} F_{n_j}(x)| \qquad (1)$$

where $Fn_j(x)$, for each $j = 1,..,k$, is the empirical distribution function (EDF) of the population j,

$$Fn_j(x) = \#\{x_i \le x\}/n_j,$$

$\mathbf{n} = (n_1,...,n_k)$, and the coefficients $c_j(\mathbf{n},x)$ satisfy

$$\Sigma_j\, c_j(\mathbf{n},x)\sqrt{n_j} = 0 \quad \text{for all } \mathbf{n} \text{ and } x,$$

and

$$\sup_n \sup_x |c_j(\mathbf{n},x)| < \infty \quad \text{for all } j = 1, ..., k.$$

It was shown (see Csörgő and Révész, 1981) that the Kiefer statistic can be approximated strongly to a Brownian bridge and a Kiefer process (See Appendix II for the definitions). This is also called the *strong embedding* property. More precisely, the following theorem holds:

THEOREM A *Let $\mathfrak{Z}_n$ be the empirical process defined above and $k > 1$. Then, there exist k independent sequences of Brownian bridges $\{B_{n_j,j}\} = \{B_{n_j}\}$ and k independent Kiefer process K_j, $j = 1,2,..,k$, such that, almost surely,*

$$\sup_x |\mathfrak{Z}_n(x) - \Sigma_j\, c_j(\mathbf{n},x) B_{n_j}(F(x))| = O(\max_j(\ln n_j/\sqrt{n_j})$$

and (2)

$$\sup_x |\mathfrak{Z}_n(x) - \Sigma_j\, c_j(\mathbf{n},x) K_j(F(x),n_j)/\sqrt{n_j}| = O(\max_j(\ln^2 n_j/\sqrt{n_j}))$$

When $k = 1$, the above is also true if the constant $c_1(n_1,x) = 1$.

The first equation implies that the empirical process $\mathfrak{Z}_n$ can be approximated by a sequence of Brownian bridges. The second equation means that the sequence of Brownian bridges converges to a Brownian bridge. More precisely, it implies (see Csörgő and Révész, 1981) that there is a Brownian bridge B(.) such that

$$P\{ \lim_n \sup_x |\mathfrak{Z}_n(x)| < \lambda \} = P\{ \sup_x |B(F(x)| < \lambda \} = P\{ \sup_t |B(t)| < \lambda \}$$

This also means that the asymptotic distribution of the empirical process is distribution free, i.e. it does not depend on F(.).

However, the Kiefer test of homogeneity in (1) has two problems in practice: (i) the test statistics depends on the selection of the coefficients; and (ii) there is no universal set of coefficients that will work for all distributions.

Another natural generalization of the K-S statistic to a k sample case is the Birnbaum-Hall test which is defined by:

$$H_n = \max\{\, |Fn_i(x)-Fn_j(x)| : 1 \le i,j \le k \,\} \qquad (3)$$

Since each of the term $|Fn_i(x)-Fn_j(x)|$ in (2) can be approximated by a Brownian bridge as suggested by Theorem A, the asymptotic distribution could be determined by $k(k-1)/2$ Brownian bridges. Unfortunately, the components $|Fn_i(x)-Fn_j(x)|$, $1 \le i,j \le k$, are correlated and the correlation depends on F(.). This implies that the asymptotic distribution of the test is not distribution free. This is the major drawback of the Birnbaum-Hall test.

A method to obtain a more practical test is to start with the goodness-of-fit tests:

Kolmogorov-Smirnov: $$S_n = \sup_x \left(\Sigma_j\, n_j(Fn_j(x)-F(x))^2 \right)^{\frac{1}{2}} \qquad (4)$$

and

Cramer-Von Mises: $$V_n = \Sigma_j \int (Fn_i(x) - F(x))^2 dF(x) \qquad (5)$$

and try to substitute $F(x)$ in (4) and (5) by some empirical distribution function. This approach is taken by the Statistical Analysis System (SAS) who substituted F by the *pooled* EDF Fn where

$$Fn(x) = \Sigma_j\, \alpha_j Fn_j(x)$$

and

$$\alpha_j = n_j/(n_1+\ldots+n_k), \quad \text{for } j = 1,2,\ldots,k.$$

in their procedure for nonparametric tests. In other words, the statistics used in SAS for testing homogeneity are:

$$S'_n = \sup_x |\Sigma_j n_j(Fn_j(x)-Fn(x))^2|^{\frac{1}{2}}$$

and (6)

$$V'_n = \Sigma_j \int (Fn_i(x) - Fn(x))^2 dFn(x)$$

Unfortunately, the terms $Fn_j(x)-Fn(x)$, $j = 1,2,\ldots,k$, are not independent. Without the independence, as was argued before, the asymptotic distribution cannot be determined easily and may not be distribution free.

This paper shows that, with a proper choice for the EDF, (4) and (5) can provide statistics that generalize the Kolmogorov-Smirnov, the Birnbaum-Hall and the Cramer-Von Mises statistics.

2. MAIN RESULTS

Let $Fn_j(x)$, $j = 1,2,..,k$, and F be defined as above. We define a k-dimension empirical process by

$$C_n(x) = \left(\sqrt{n_1}(Fn_1(x)-F(x)),\ldots, \sqrt{n_k}(Fn_k(x)-F(x)) \right) \qquad (7)$$

From Theorem A, it is easy to prove that the process $\mathbf{C}_n$ can be embedded strongly to a k-dimension Brownian bridge and a k-dimension Kiefer process. The exact statement of this result is the following theorem which presents a more general form of the K-S goodness-of-fit test given in (4). The proves of the mains theorems are given in Section 3.

THEOREM 1 *Let* $\mathbf{C}_n(x)$ *be defined as above. There exist a sequence of k-dimension Brownian bridges* $\{\mathbf{B}_n\}$ *and k independent Kiefer processes* $K_1,...,K_k$ *such that, almost surely,*

$$\sup_x \| \mathbf{C}_n(x) - \mathbf{B}_n(\mathbf{F}(x)) \| = O(\max_j(\ln n_j/\sqrt{n_j}))$$

and (8)

$$\sup_x \| \mathbf{C}_n(x) - \mathbf{K}(\mathbf{F}(x),\mathbf{n}) \| = O(\max_j(\ln^2 n_j/\sqrt{n_j}))$$

where $\mathbf{K}(t,\mathbf{n}) = (K_1(t,n_1)/\sqrt{n_1},...,K_k(t,n_k)/\sqrt{n_k})$, $\mathbf{n} = (n_1,..,n_k)$ *and* $\|.\|$ *is any norm of a k-dimension Euclidean space.*

Because in a finite dimensional Euclidean space all the norms are equivalent, it is sufficient to give the proofs in Euclidean distance,

$$\| \mathbf{x} \| = \sqrt{(\Sigma_i \, x_i^2)}$$

where $\mathbf{x} = (x_1,...,x_k)$. The $\mathbf{S}_n$ given in (4) and $\mathbf{V}_n$ in (5) can be obtained by taking the Euclidean distance on $\mathbf{C}_n$. In other words, we can write $\mathbf{S}_n = \sup_x \| \mathbf{C}_n(x) \|$ and similarly, $\mathbf{V}_n = \int \| \mathbf{C}_n(x) \|^2 d\mathbf{F}(x)$. Furthermore, since $\mathbf{V}_n \le \mathbf{S}_n$, Theorem 1 encompasses both the K-S and the Cramer-Von Mises goodness-of-fit tests.

To test the homogeneity hypothesis, it is necessary to define a statistic that does not depend on the distribution F. This can be obtained as follows. For each j, $1 \le j \le k$, select at random half of the n_j samples. If n_j is odd, discard one sample at random and redefine n_j as n_j-1. Then, select $\frac{1}{2}n_j$ samples at random. For convenience, the rest of the paper assumes all n_j to be even. For each $j = 1,2,...,k$, define

$\mathbf{F}_{n_j/2}(x)$ = EDF constructed from the randomly selected $n_j/2$ samples

and

$\mathbf{F}_{\bar{n}_j/2}(x)$ = EDF constructed the remaining $n_j/2$ samples

Then, it is easy to see that

$$\mathbf{F}_{n_j}(x) = (\mathbf{F}_{n_j/2}(x) + \mathbf{F}_{\bar{n}_j/2}(x))/2$$

Now, define the *half pooled* EDF $\mathbf{F}_{n/2}(x)$ by

$$\mathbf{F}_{n/2}(x) = \Sigma_j \, \alpha_j \mathbf{F}_{n_j/2}(x)$$

where

$$\alpha_j = n_j/(n_1+...+n_k), \text{ for } j = 1,2,...,k.$$

The EDF $F_{n/2}(x)$ is equal to the EDF constructed from all the selected half samples. The main result is given by the following theorem that is based on the k-dimension process

$$\mathbb{C}_n(x) = (\sqrt{n_1}(F_{n_1}(x)-F_{n/2}(x)),..., \sqrt{n_k}(F_{n_k}(x)-F_{n/2}(x))). \tag{9}$$

The above k-dimension process is in fact a k-vector valued process. We call a k-vector valued process a *k-dimension Brownian bridge* if all the components of this process are Brownian bridges and are also independent. The independence of the components is essential to our results.

THEOREM 2 *Given the above, there exist a sequence of k-dimension Brownian bridges $\{\mathbf{B}_n\}$ such that, with probability one,*

$$\sup_x \| \mathbb{C}_n(x) - \mathbf{B}_n(F(x)) \| = O(\max_j(\ln(n_j/2)/(n_j/2)^{\frac{1}{2}})) \tag{10}$$

where and $\|.\|$ is any norm of a k-dimension Euclidean space.

The importance of embedding to a Kiefer process is to ensure that the Brownian bridges converge and that the limiting process is also a Brownian bridge. To state an analogous embedding result to a k-dimensional Kiefer process, the sample sizes are restricted to fixed ratios, i.e.

$$n_i = \alpha_i n$$

with

$$\Sigma_i \alpha_i = 1.$$

THEOREM 3 *Let n_i and α_i, $i = 1,2,..,k$, be defined as above. Then, there exist k independent Kiefer processes $K_1,...,K_k$ such that, with probability one,*

$$\sup_x \| \mathbb{C}_n(x) - \mathbf{K}(F(x),n) \| = O(\max_j \ln^2(n_j/2)/(n_j/2)^{\frac{1}{2}})) \tag{11}$$

where $\mathbf{K}(t,n) = (K_1(t,n_1)/\sqrt{n_1},...,K_k(t,n_k)/\sqrt{n_k})$ and $\|.\|$ is any norm of a k-dimension Euclidean space.

Remark: In Section 4, it will be discussed that the restriction of fixed ratios can be relaxed. However, without this restriction, the statement of Theorem 3 will be more wordy.

Although all Euclidean norms are equivalent, different norms will give different statistics for the test of homogeneity as stated by the following corollaries:

Corollary 1 *If* $\|.\|$ *is the Euclidean distance, then the statistic*

$$\mathfrak{C}_a = \sup_x |\Sigma_j\, n_j(F_{n_j}(x)-F_{n/2}(x))^2|^{\frac{1}{2}}$$

generalize the Kolmogorov-Smirnov statistic of (4). *The asymptotic distribution of this test can be determined by k independent Brownian bridges.*

Corollary 2 *If* $\|.\|$ *is the maximum norm, i.e.*

$$\|x\| = \max(\,|x_j| : j = 1,2,..,k\,)$$

then

$$\mathfrak{C}_b = \sup_x \max_j \sqrt{n_j}|F_{n_j}(x)-F_{n/2}(x)|$$

is a test statistic that generalize the Birnbaum-Hall test.

Remark: In the statistic $\mathfrak{C}_b$, since the corresponding Brownian bridges are independent, the distribution of the statistic in Corollary 2 is much easier to obtain than that of the Birnbaum-Hall statistics. Also, the K-S 2-sample test is equivalent to Corollaries 1 and 2.

The Cramer-Von Mises statistic is based on the $\mathcal{L}_2$ norm of the functions. Since the $\mathcal{L}_2$ norm is bounded by the supremum norm, the Cramer-Von Mises test can be generalized:

Corollary 3 *Let* x_m *denote the observations used in the calculation of* $F_{n/2}(x)$. *Then the test statistic*

$$\mathfrak{C}_c = \Sigma_j\, \Sigma_m\, n_j|F_{n_j}(x_m)-F_{n/2}(x_m)|^2$$

is a generalization of the Cramer-Von Mises test of goodness-of-fit. Furthermore, both tests have the same asymptotic distribution.

3. PROOF OF THE THEOREMS

The proof of Theorem 1 is a trivial application of Theorem A. The following lemma is needed to prove Theorems 2 and 3.

Lemma *Let* x_{ij}, $j = 1,2,..,n_i$ *and* $i = 1,2$, *be two sets of random independent observations of two random variables. And, let* F_{n_1} *and* F_{n_2} *be the their respective* EDFs. *Then,*

$$E[F_{n_1}(x)F_{n_2}(y)] = F(x)F(y)$$

and, for $i = 1$ or 2,

$$E[F_{n_i}(x)F_{n_i}(y)] = F(x)F(y) - \{F(x\wedge y) - F(x)F(y)\}/n_i$$

where $x\wedge y$ denotes the minimum of x and y

Proof Note that F_{n_i} can be expressed by sums of characteristic functions; i.e., for i= 1 or 2,

$$F_{n_i}(x) = (\Sigma_j I_x(x_j))/n_i \tag{12}$$

where $I_x(u) = 1$ or 0 if $u \le x$ or $u > x$, respectively. Since $E[I_x(x_i)] = F(x)$, and

$$E[(I_x(x_{ij})(I_y(x_{im})] = F(x \wedge y) \qquad (\text{for } i = 1 \text{ or } 2)$$

and

$$E[(I_x(x_{1i})(I_y(x_{2j})] = F(x)F(y)$$

for any j and m, from (12), we have

$$E[F_{n_1}(x)F_{n_2}(y)] = \Sigma_{j,m} E[I_x(x_{1j})I_x(x_{2m})]/n_1n_2 = F(x)F(y)$$

and

$$E[F_{n_i}(x)F_{n_i}(y)] = \Sigma_{j \neq m} E[I_x(x_{ij})I_x(x_{im})]/n_i^2 + \Sigma_j E[I_x(x_{ij})I_x(x_{ij})]/n_i^2$$

$$= F(x)F(y) - (F(x \wedge y) - F(x)F(y))/n_i$$

This proves the lemma.

Proof of Theorem 2.

To prove Theorem 2, we must show that (i) the process $\mathfrak{C}_n(x)$ has a diagonal covariance matrix, and (ii) we can construct a sequence of k-dimension Brownian bridge which has the same covariance matrix and satisfies the conditions in (10).

To show that the k-dimension empirical process has a diagonal covariance matrix, let $F_{n_j}(x)$, j = 1,2,..,k, and $F_{n/2}(x)$ be defined as above. The following equalities are direct consequences of the above Lemma:

$$E[F_{n_i}(x)F_{n_j}(x)] = \begin{cases} F(x)F(y) + (F(x \wedge y) - F(x)F(y))/n_i & \text{if } i = j \\ F(x)F(y) & \text{otherwise} \end{cases} \tag{13}$$

$$E[(F_{n_i}(x)F_{n/2}(y))] = F(x)F(y) + (F(x \wedge y) - F(x)F(y))/n \tag{14}$$

$$= E[F_{n/2}(x)F_{n_j}(y)]$$

$$E[F_{n/2}(x)F_{n/2}(y))] = F(x)F(y) + (F(x \wedge y) - F(x)F(y))/(n/2) \tag{15}$$

where $n = n_1+...+n_k$. From (13) to (15), it follows easily that

$$E[(F_{n_i}(x)-F_{n/2}(x))(F_{n_i}(y)-F_{n/2}(y))] = \begin{cases} (F(x \wedge y) - F(x)F(y))/n_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

The proof of Theorem 2 will be completed by constructing a sequence of k-dimension Brownian bridges which approximates the process $\mathbb{G}_n(x)$ and whose covariance matrix is the same as that of $\mathbb{G}_n(x)$. Let $F_{n_i/2}$ and $\tilde{F}_{n_i/2}$ be defined as above. Since $\Sigma_i\ \alpha_i=1$, we can write

$$\sqrt{n_i}(F_{n_i}(x)-F_{n/2}(x)) = \tfrac{1}{2}\sqrt{2}\{\sqrt{(n_i/2)}(\tilde{F}_{n_i/2}(x)-F(x))\} + \tfrac{1}{2}\sqrt{2}\{\sqrt{(n_i/2)}(F_{n_i/2}(x)-F(x))\}$$

$$- \Sigma_j\ \alpha_j\sqrt{(2n_i/n_j)}\{\sqrt{(n_j/2)}(F_{n_j/2}(x)-F(x))\}$$

From Theorem 1, there exist a sequence of $2k$-dimension Brownian bridges

$$\{\ (\ B\tilde{n}_1,..,\ B\tilde{n}_k, Bn_1^*,..,Bn_k^*\ )\ \}_n$$

such that each i, and almost surely,

$$\begin{aligned} &\sup_x |(\sqrt{n_i}/2)(F_{n_i/2}(x)-F(x))-B_{n_i}^*(F(x))| = O(\max_j(\ln(n_j/2)/\sqrt{(n_j/2)}), \\ &\sup_x |(\sqrt{n_i}/2)(\tilde{F}_{n_i/2}(x)-F(x))-B\tilde{n}_i(F(x))| = O(\max_j(\ln(n_j/2)/\sqrt{(n_j/2)}), \end{aligned} \tag{16}$$

Define, for $0 \le t \le 1$,

$$B_{n_i}(t) = \tfrac{1}{2}\sqrt{2}B\tilde{n}_i(t) + \tfrac{1}{2}\sqrt{2}Bn_i^*(t) - \Sigma_j\ (\alpha_j\sqrt{(2n_i/n_j)}Bn_j^*(t)$$

Then, from the above definition of B_{n_i} and from the fact that $B\tilde{n}_i$ and Bn_i^* are independent Brownian bridges, it is easy to show that

$$E[B_{n_i}(s)B_{n_j}(t)] = \begin{cases} s\wedge t - st & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Define a k vector of Brownian bridge as follow:

$$\mathbf{B}_n = (\ B_{n_1},...,B_{n_k}\ )$$

Equations (17) shows that $\mathbf{B}_n$ is a k-dimension Brownian bridge. Also, from (16), it follows easily that they satisfy condition (10) of the Theorem 2.

Proof of Theorem 3.

From Theorem 1, there exist $2k$ independent Kiefer processes

$$K_1^*,...,K_k^*,K_1^{\sim},...,K_k^{\sim}$$

such that each i, and almost surely,

$$\sup_x |(\sqrt{n_i}/2)(F_{n_i/2}(x)-F(x))-K_i^*(F(x),n_i/2)/\sqrt{(n_i/2)}| = O(\max_j(\ln^2(n_j/2)/\sqrt{(n_j/2)}))$$
$$\sup_x |(\sqrt{n_i}/2)(F_{\bar{n}_i/2}(x)-F(x))-K_i^{\sim}(F(x),n_i/2)/\sqrt{(n_i/2)}| = O(\max_j(\ln^2(n_j/2)/\sqrt{(n_j/2)})). \quad (18)$$

Now, given that $n_i = \alpha_i n$ and $\Sigma_i \alpha_i = 1$, define:

$$K_i(t,n) = K_i^{\sim}(t,n_i/2) + K_i^*(t,n_i/2) - 2\alpha_i \Sigma_j K_j^*(t,n_j/2)$$

for $0 \le t \le 1$. From (17), it follows that K_i's satisfy condition (11) of Theorem 3.

To complete the proof, it is necessary to proof that they are also independent, ie., it is necessary to show that

$$E[K_i(s,n)K_j(t,m)] = \begin{cases} \alpha_i(s\wedge t - st)(n\wedge m) & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

But this follows easily from the definition of K_i's.

4. DISCUSSION

To understand the problems with the statistics used by SAS (equations 6), it is observed that the result of the Lemma implies that, for all $i \neq j$,

$$E[\sqrt{n_i}(F_{n_i}(x)-F_n(x))\sqrt{n_j}(F_{n_j}(y)-F_n(y))] = -(F(x\wedge y)-F(x)F(y))(\sqrt{(n_i n_j)}/n)$$

where n is the sum of all n_i's. As n approaches to ∞, the right hand side term of the above equation does not vanish. This means that components of the limiting process of S'_n of (6) have correlations depending on F. Therefore, the asymptotic distribution of the test may also depend on F, ie, it is not a distribution free test. Also, note that the presence of the correlation implies that the variance of S'_n is different from the sum of the variances of the components.

As was mentioned in Section 2, it is possible to prove a somewhat more general version of Theorem 3. More precisely, we can show that there exist k $(k+1)$-parameter Gaussian processes, $K_i(t,\mathbf{n})$, $\mathbf{n} = (n_1,\ldots,n_k)$, $k = 1,2,\ldots,k$, not necessary independent, such that (11) of Theorem 3 is satisfied. These Gaussian processes have zero means and the following covariance structure: for $i \neq j$ and $\mathbf{m} \neq \mathbf{n}$

$$E[K_i(s,\mathbf{n})K_i(t,\mathbf{m})] = \phi(\mathbf{n}\wedge\mathbf{m})(s\wedge t - st)/2,$$

$$E[K_i(s,\mathbf{n})K_i(t,\mathbf{n})] = n_i(s\wedge t - st)/2$$

and

$$E[K_i(s,\mathbf{n})K_j(t,\mathbf{n})] = 0$$

where ϕ is a real valued k variable function and $\mathbf{n}\wedge\mathbf{m} = (n_1\wedge m_1,\ldots,n_k\wedge m_k)$. These processes behave like Kiefer processes only when $\mathbf{n}$ is fixed. Also, when $\mathbf{n}$ is fixed, they are independent. Furthermore, these Gaussian process can be constructed from $2k$ Kiefer process. This can assure that the limiting process is a Brownian bridge.

To make the procedures given in this paper more practical, it will be necessary to study the power of the tests based on the process $\mathbb{C}_{\mathbf{n}}(x)$ and the sensitivity of the tests for different subsample selections.

APPENDIX II

DEFINITIONS

A **Gaussian process** $X(t)$ *is a continuous stochastic process such that for any* $t_1,\ldots,t_n$, *the joint distribution of* $X(t_1),\ldots,X(t_n)$ *is multi-variate normal. The process* $X(t)$ *is completely determined by its covariance structure:* $\mathrm{E}[X(t_i)X(t_j)]$.

A **Brownian bridge** $\mathrm{B}(t)$, $0 \le t \le 1$, *is a Gaussian process with zero mean and covariance structure:*

$$\mathrm{E}[\mathrm{B}(s)\mathrm{B}(t)] = s \wedge t - st$$

for all $0 \le s, t \le 1$.

A **Kiefer process** $K(t,n)$, $0 \le t \le 1$ *and* $n \ge 0$, *is a 2-parameter Gaussian process with zero mean and covariance structure: for some constant* $a > 0$,

$$\mathrm{E}[K(t,n)K(s,m)] = a(s \wedge t - st)(n \wedge m)$$

for all $0 \le s, t \le 1$ *and* $n, m \ge 0$.

A **k-dimension Brownian bridge** $\mathbf{B}(t) = (B_1(t),\ldots,B_k(t))$, $0 \le t \le 1$, *is a vector of k independent brownian bridges with the following covariance structure:*

$$\mathrm{E}[B_i(s)B_j(t)] = \begin{cases} 0 & \text{if } i \neq j \\ s \wedge t - st & \text{if } i = j \end{cases}$$

for any pair of s and t such that $0 \le s, t \le 1$.

A **k-dimension Kiefer process** $\mathbf{K}(t,n) = (K_1(t,n),\ldots,K_k(t,n))$, *is a vector of k independent Kiefer processes with covariance structure:*

$$\mathrm{E}[K_i(s,n)K_j(t,m)] = \begin{cases} 0 & \text{if } i \neq j \\ a_i(s \wedge t - st)(n \wedge m) & \text{if } i = j \end{cases}$$

REFERENCE

[1] Kolmogorov, A.N., (1931), "Sulla determinazione empirica di une legge di distribuzione", Giorn. Inst. Ital. Attuari **4**, 83-91.

[2] Smirnov, N.V., (1939), "On the estimation of discrepancy between empirical curves of distribution for two independent samples", Bull. Math. de l'Université de Moscou **2** (Fasc.2).

[3] Kiefer, J., (1959) "*K*-sample analogues of the Kolmogorov-Smirnov and Cramer-von Mises tests", Ann. Math. Statistics, **30**, 321-332.

[4] Burke, M.D. and Csörgő, M., (1976) "Strong approximations of the *k*-sample quantile and multi-variate empirical process", Carleton Mathematical Lecture Notes. No. 16.

[5] Birnbaum, Z.W. and Hall, R.A., (1960), "Small sample distributions for multi-sample statistics of the Smirnov type", Ann. of Math. Statistics, **22**, 592-596.

[6] Csörgő, M. and Révész, P., (1981), *Strong Approximations in Probability and Statistics*, Akadémiai Kiadó, Budapest.